Sparsity Patterns

Michael Komodromos

May 4, 2022

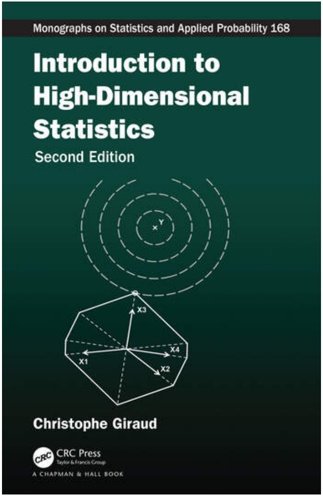
Outline

About this presentation

About this presentation

Based on the 2nd Chapter of “An introduction to High-dimensional Statistics” by Giraud (2021).

Can be found online for free



Motivation: Why should we care?

Motivation: Why should we care?

In many applications we work with high-dimensional data.

Motivation: Why should we care?

In many applications we work with high-dimensional data.

Making use of the data requires tools that take into account the patterns within it.

Motivation: Why should we care?

In many applications we work with high-dimensional data.

Making use of the data requires tools that take into account the patterns within it.

The aim of this presentation is to introduce some of these patterns and the corresponding tools to analyze them

Statistical Setting

Sparsity patterns: Statistical setting

Sparsity patterns: Statistical setting

Before we jump the different types of sparsity, we need to layout the statistical setting

Sparsity patterns: Statistical setting

Before we jump the different types of sparsity, we need to layout the statistical setting

Statistical setting

Formally, we'll be working in a regression setting where,

$$y_i = f(x_i) + \epsilon_i \tag{1}$$

which links our response $y \in \mathbb{R}$, to p variables stored in a p-dimensional real valued vector $x_i \in \mathbb{R}^p$.

Main focus

Main focus

Our main focus is going to be on linear regression models where

$$f(x) = \sum_{j \in J} \beta_j x_j \tag{2}$$

Main focus

Our main focus is going to be on linear regression models where

$$f(x) = \sum_{j \in J} \beta_j x_j \tag{2}$$

This is because many forms for $f(x)$ can be re-written in this way, e.g. piecewise constant regression, additive models etc.

Sparsity Patterns

Types of Sparsity

Types of Sparsity

There are three main sparsity patterns we'll discuss:

Types of Sparsity

There are three main sparsity patterns we'll discuss:

- Coordinate sparsity (which we've all come across)

Types of Sparsity

There are three main sparsity patterns we'll discuss:

- Coordinate sparsity (which we've all come across)
- Group-wise sparsity

Types of Sparsity

There are three main sparsity patterns we'll discuss:

- Coordinate sparsity (which we've all come across)
- Group-wise sparsity
- Sparse group-wise sparsity

Sparsity patterns: Coordinate sparsity

Sparsity patterns: Coordinate sparsity

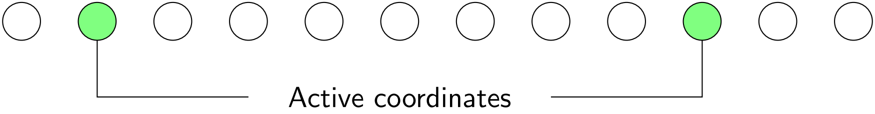
Coordinate sparsity

Only a few coordinates of β are non-zero

Sparsity patterns: Coordinate sparsity

Coordinate sparsity

Only a few coordinates of β are non-zero



Sparsity patterns: Group sparsity

Sparsity patterns: Group sparsity

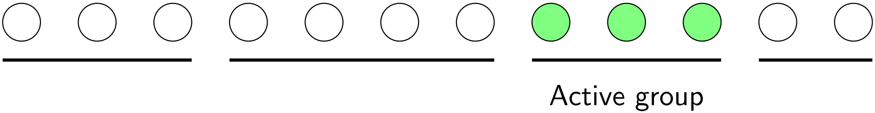
Group sparsity

The coordinates of β are clustered into groups, and only few of those groups are non-zero.

Sparsity patterns: Group sparsity

Group sparsity

The coordinates of β are clustered into groups, and only few of those groups are non-zero.



Sparsity patterns: Sparse-group sparsity

Sparsity patterns: Sparse-group sparsity

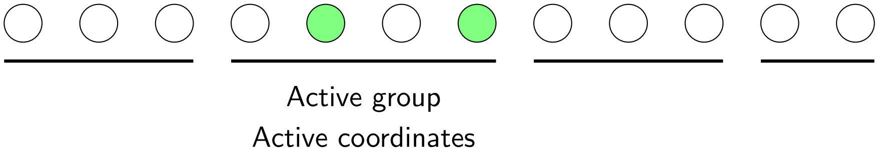
Sparse-group sparsity

The coordinates of β are clustered into groups, and only few of the coordinates within a group are non-zero.

Sparsity patterns: Sparse-group sparsity

Sparse-group sparsity

The coordinates of β are clustered into groups, and only few of the coordinates within a group are non-zero.



Examples in Practice

These examples are taken from

"A Sparse-group LASSO" by Simon et al. Sec 5.

A SPARSE-GROUP LASSO

NOAH SIMON, JEROME FRIEDMAN, TREVOR HASTIE, AND ROB TIBSHIRANI

Abstract. For high dimensional supervised learning problems, often using problem specific assumptions can lead to greater accuracy. For problems with grouped covariates, which are believed to have sparse effects both on a group and within group level, we introduce a regularized model for linear regression with ℓ_1 and ℓ_2 penalties. We discuss the sparsity and other regularization properties of the optimal fit for this model, and show that it has the desired effect of group-wise and within group sparsity. We propose an algorithm to fit the model via accelerated generalized gradient descent, and extend this model and algorithm to convex loss functions. We also demonstrate the efficacy of our model and the efficiency of our algorithm on simulated data.

Keywords: penalize, regularize, regression, model, nesterov

Example: Breast cancer data

Example: Breast cancer data

Dataset

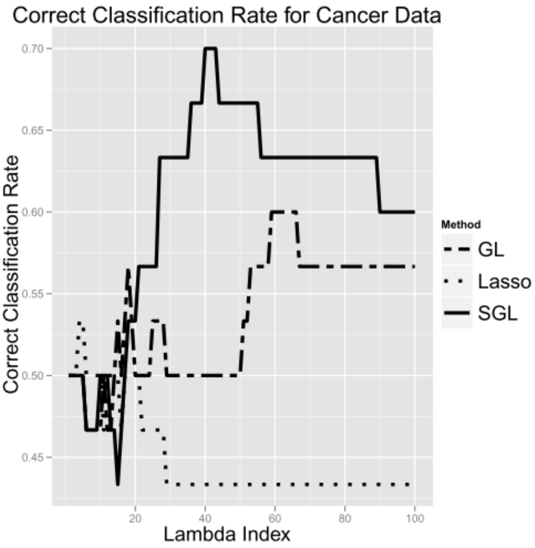
Example: Breast cancer data

Dataset

- Gene expression values of $n = 60$ patients with estrogen positive breasts
- Patients were treated with tamoxifen for 5 years
- Classified according to whether cancer recurred
- After pre-processing $p \approx 12,000$ genes
- Genes are groups by cytogenetic position data (GSEA C1 data)
- 30 patients chosen at random used in the training set.

Example: Breast cancer data

Example: Breast cancer data



Example: Breast cancer data

Example: Breast cancer data

Examining the peak classification accuracy for each method

Example: Breast cancer data

Examining the peak classification accuracy for each method

Method	Classification Accuracy	Num. Features
Sparse group LASSO	70%	54 (11 groups)
Group LASSO	60%	74 (14 groups)
LASSO	53%	3

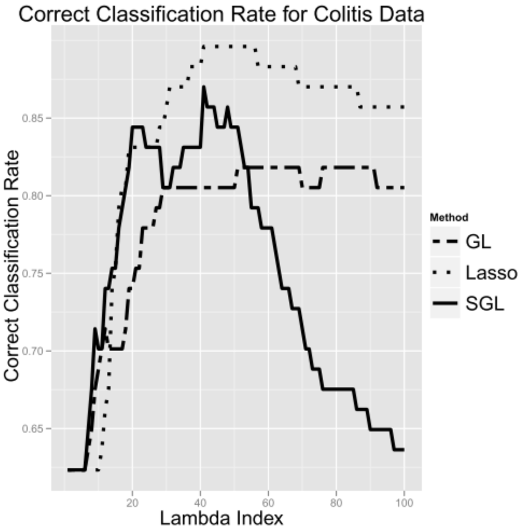
Example: Colitis data

Example: Colitis data

- 127 patients
- 85 with colitis, 42 controls
- $p \approx 8,300$ after pre-processing
- grouped into gene-sets using cytogenetic information giving 277 disjoint groups.
- 50 observations used to fit models, 77 used to test.

Example: Colitis data

Example: Colitis data



Example: Colitis data

Example: Colitis data

Examining the peak classification accuracy for each method

Example: Colitis data

Examining the peak classification accuracy for each method

Method	Classification Accuracy	Num. Features
LASSO	90%	19
Sparse group LASSO	87%	43 (8 groups)
Group LASSO	84%	36 (7 groups)

Example: Colitis data

Examining the peak classification accuracy for each method

Method	Classification Accuracy	Num. Features
LASSO	90%	19
Sparse group LASSO	87%	43 (8 groups)
Group LASSO	84%	36 (7 groups)

In the second example the sparse group LASSO with the chosen gene sets did not perform as well as the LASSO - specialist information about the gene set may improve the results.

Interim

Interim

Take home message

There are many different sparsity patterns for regression models.

Some may be more suited for solving problems than others.

Interim

Take home message

There are many different sparsity patterns for regression models.

Some may be more suited for solving problems than others.

When to use which? A rule of thumb

Method	Number of Groups	Size of groups
Coordinate sparse	Small	Large
Group sparse	Large	Small
Sparse group sparse	Large	Large

Questions so far?

We're about to jump into some theory

so now's the time to delay that happening!

Model selection

Model selection

Model selection

General idea

Compare different statistical models corresponding to different possible structures and select the model best suited to estimation.

Model selection: Framework

Model selection: Framework

Framework

We're assuming

$$y_i = f_i + \epsilon_i, \quad \epsilon_i \overset{iid}{\sim} N(0, \sigma^2) \tag{3}$$

where $f_i = \langle x_i, \beta \rangle$, with $\beta = (\beta_1, \ldots, \beta_p)^\top$ and $x_i = (x_{i,1}, \ldots, x_{i,p}) \in \mathbb{R}^p$ for $i = 1, \ldots, n$.

We will assume a true and unknown $\beta^* \in \mathbb{R}^p$, used to generate the data.

Model selection: Framework

Model selection: Framework

Our aim is to provide an estimate for f_i that bests recovers $f_i^* = \langle x_i, \beta^* \rangle$.

Model selection: Framework

Our aim is to provide an estimate for f_i that bests recovers $f_i^* = \langle x_i, \beta^* \rangle$.

To keep things simple we're going to be considering the model selection process for co-ordinate sparsity.

Model selection

Model selection

Model selection under co-ordinate sparsity involves:

Model selection

Model selection under co-ordinate sparsity involves:

1. Considering the collection of models $\{S_m : m \in \mathcal{M}\}$ where $S_m = \text{span}\{X_j : j \in m\}$ and $\mathcal{M}$ is a set of all subsets of $\{1, \ldots, p\}$.

Model selection

Model selection under co-ordinate sparsity involves:

1. Considering the collection of models $\{S_m : m \in \mathcal{M}\}$ where $S_m = \text{span}\{X_j : j \in m\}$ and $\mathcal{M}$ is a set of all subsets of $\{1, \ldots, p\}$.

2. Estimating the model $\widehat{f}_m$ for each m using maximum likelihood.

Model selection

Model selection under co-ordinate sparsity involves:

1. Considering the collection of models $\{S_m : m \in \mathcal{M}\}$ where $S_m = \text{span}\{X_j : j \in m\}$ and $\mathcal{M}$ is a set of all subsets of $\{1, \ldots, p\}$.
2. Estimating the model $\widehat{f}_m$ for each m using maximum likelihood.
3. Estimating f by selecting the **best** model $\widehat{f}_m$ from the collection $m \in \mathcal{M}$

Model selection

Model selection

To quantify what we mean by **best**, we will use the ℓ^2 risk,

Model selection

To quantify what we mean by **best**, we will use the ℓ^2 risk,

$$r_m = \mathbb{E}\left[\|\widehat{f}_m - f^*\|^2\right] \tag{4}$$

Model selection

To quantify what we mean by **best**, we will use the ℓ^2 risk,

$$r_m = \mathbb{E}\left[\|\widehat{f}_m - f^*\|^2\right] \tag{4}$$

As we don't know f^* a natural idea is to estimate r_m using an unbiased estimator, and since,

$$r_m = \mathbb{E}\left[\|Y - f_m\|^2\right] + (2d_m - n)\sigma^2 \tag{5}$$

where $d_m = \dim(S_m)$, we will consider the estimator

Model selection

To quantify what we mean by **best**, we will use the ℓ^2 risk,

$$r_m = \mathbb{E}\left[\|\widehat{f}_m - f^*\|^2\right] \tag{4}$$

As we don't know f^* a natural idea is to estimate r_m using an unbiased estimator, and since,

$$r_m = \mathbb{E}\left[\|Y - f_m\|^2\right] + (2d_m - n)\sigma^2 \tag{5}$$

where $d_m = \dim(S_m)$, we will consider the estimator

$$\widehat{r}_m = \|Y - f_m\|^2 + (2d_m - n)\sigma^2 \tag{6}$$

AIC

AIC

In turn we can estimate m using

$$\widehat{m} \in \underset{m \in \mathcal{M}}{\operatorname{argmin}} \left\{ \| Y - f_m \|^2 \; + 2 d_m \sigma^2 \right\} \tag{7}$$

Where we have dropped the $-n\sigma^2$ term (which does not change $\widehat{m}$). This estimator for $\widehat{m}$ is the popular Akaike information criterion.

Penalization

Penalization

However, model selection based on $\widehat{m}$ can produce poor results in practise.

Penalization

However, model selection based on $\widehat{m}$ can produce poor results in practise.

This is because, for a large number of possible models $\widehat{r}_m$ will deviate substantially from the mean, hence $\widehat{m}$ will favour larger models.

Penalization

However, model selection based on $\widehat{m}$ can produce poor results in practise.

This is because, for a large number of possible models $\widehat{r}_m$ will deviate substantially from the mean, hence $\widehat{m}$ will favour larger models.

To correct this, we can introduce a penalization term taking into account the number of models per dimension.

Penalization

Penalization

One such penalization scheme (motivated by theory, which I've skipped) is given by

Penalization

One such penalization scheme (motivated by theory, which I've skipped) is given by

$$\widehat{m} \in \underset{m \in \mathcal{M}}{\operatorname{argmin}} \left\{ \| Y - f_m \|^2 \ + \lambda |m| \right\} \tag{8}$$

where $|\cdot|$ is the cardinality of the set m and $\lambda = (1 + \sqrt{2 \log p})^2 \sigma^2$.

Penalization

One such penalization scheme (motivated by theory, which I've skipped) is given by

$$\widehat{m} \in \underset{m \in \mathcal{M}}{\operatorname{argmin}} \left\{ \|Y - f_m\|^2 + \lambda |m| \right\} \tag{8}$$

where $|\cdot|$ is the cardinality of the set m and $\lambda = (1 + \sqrt{2 \log p})^2 \sigma^2$.

Theory is covered at the end of Chapter 2 and start of Chapter 5 for Giraud (2021).

Penalization

Penalization

Importantly the previous formulation can be re-written as,

Penalization

Importantly the previous formulation can be re-written as,

$$\widehat{m} \in \underset{m \in \mathcal{M}}{\operatorname{argmin}} \min_{\beta : \mathrm{supp}(\beta) = m} \left\{ \|Y - f_m\|^2 + \lambda |\beta|_0 \right\} \tag{9}$$

where $|\beta|_0 = \#\{\beta_j \neq 0 : j = 1, \ldots, p\}$

Penalization: LASSO

Penalization: LASSO

Solving

$$\widehat{\beta}_{\lambda} \in \underset{\beta:\mathrm{supp}(\beta)=m}{\operatorname{argmin}} \left\{ \|Y - f_m\|^2 + \lambda|\beta|_0 \right\} \tag{10}$$

is non-convex due to $|\beta|_0$.

Penalization: LASSO

Solving

$$\widehat{\beta}_\lambda \in \underset{\beta:\text{supp}(\beta)=m}{\text{argmin}} \left\{ \| Y - f_m \|^2 \ + \lambda |\beta|_0 \right\} \tag{10}$$

is non-convex due to $|\beta|_0$.

Relaxing $|\beta|_0$ to be $|\beta|_1$ is convex and gives us the popular LASSO estimator

Recap

Recap

How we got here:

- We wanted to do model selection

Recap

How we got here:

- We wanted to do model selection
- But the estimator $\widehat{m}$ when p is large tends to not work well.

Recap

How we got here:

- We wanted to do model selection
- But the estimator $\widehat{m}$ when p is large tends to not work well.
- Penalizing the estimator has better (statistical) proprieties but is computationally intractable

Recap

How we got here:

- We wanted to do model selection
- But the estimator $\widehat{m}$ when p is large tends to not work well.
- Penalizing the estimator has better (statistical) proprieties but is computationally intractable
- Relaxing our penalization scheme is computationally tractable and has decent statistical proprieties

A Bayesian perspective

Some history

Some history

Model selection priors started gaining attention around the 1990s

Bayesian Variable Selection in Linear Regression

T. J. MITCHELL and J. J. BEAUCHAMP*

This article is concerned with the selection of subsets of predictor variables in a linear regression model for the prediction of a dependent variable. It is based on a Bayesian approach, intended to be as objective as possible. A probability distribution is first assigned to the dependent variable through the specification of a family of prior distributions for the unknown parameters in the regression model. The method is not fully Bayesian, however, because the ultimate choice of prior distribution from

Recent work has been aimed at scaling these prior to suit high-dimensional problems.

Variable Selection Via Gibbs Sampling

EDWARD I. GEORGE and ROBERT E. McCULLOCH*

A crucial problem in building a multiple regression model is the selection of predictors to include. The main thrust of this article is to propose and develop a procedure that uses probabilistic considerations for selecting promising subsets. This procedure entails embedding the regression setup in a hierarchical normal mixture model where latent variables are used to identify subset choices. In this framework the promising subsets of predictors can be identified as those with higher posterior probability. The computational burden is then alleviated by using the Gibbs sampler to indirectly sample from this multinomial posterior distribution on the set of possible subset choices. Those subsets with higher probability—the promising ones—can then be identified by their more frequent appearance in the Gibbs sample.

Bayesian model selection

Bayesian model selection

The idea: place a prior over the different possible models $m \in \mathcal{M}$.

Bayesian model selection

The idea: place a prior over the different possible models $m \in \mathcal{M}$.

Construct a posterior that updates our prior belief about model m using the data.

Bayesian model selection

The idea: place a prior over the different possible models $m \in \mathcal{M}$.

Construct a posterior that updates our prior belief about model m using the data.

This is a (conceptually) rich tool, as it assigns a posterior probability to each model m and therefore gives us a nice way to do model and variable selection.

Bayesian model selection: Prior

Bayesian model selection: Prior

Co-ordinate sparse priors

Bayesian model selection: Prior

Co-ordinate sparse priors

$$\begin{aligned} \beta_j | z_j &\overset{ind}{\sim} z_j \Phi(\beta_j) + (1 - z_j)\delta_0 \\ z_j &\overset{iid}{\sim} \text{Bernoulli}(p) \end{aligned} \tag{11}$$

for $j = 1, \ldots, p$ and where Φ is a continuous distribution for β and δ_0 is a Dirac mass at 0.

Bayesian model selection: Prior

Bayesian model selection: Prior

Group-wise and Sparse-Group sparsity

Bayesian model selection: Prior

Group-wise and Sparse-Group sparsity

$$\begin{aligned} \beta_{G_k} | z_k &\overset{ind}{\sim} z_k \Phi(\beta_{G_k}) + (1 - z_k)\delta_0(\beta_{G_k}) \\ z_k &\overset{iid}{\sim} \text{Bernoulli}(p) \end{aligned} \tag{12}$$

where $G_k \subset \{1, \ldots, p\}$ for $k = 1, \ldots, m$ are disjoint sets such that $\cup_{k=1}^{m} G_k = \{1, \ldots, p\}$ and δ_0 is a multivariate Dirac mass at 0.

Bayesian model selection: Prior

Group-wise and Sparse-Group sparsity

$$\begin{aligned} \beta_{G_k} | z_k &\overset{ind}{\sim} z_k \Phi(\beta_{G_k}) + (1 - z_k)\delta_0(\beta_{G_k}) \\ z_k &\overset{iid}{\sim} \text{Bernoulli}(p) \end{aligned} \tag{12}$$

where $G_k \subset \{1, \ldots, p\}$ for $k = 1, \ldots, m$ are disjoint sets such that $\cup_{k=1}^m G_k = \{1, \ldots, p\}$ and δ_0 is a multivariate Dirac mass at 0.

Sparse-group sparsity can be achieved by ensuring Φ induces sparsity e.g. using another spike and slab prior.

Computational problems

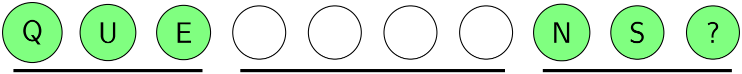
Computational problems

The same issues encountered earlier are experienced, namely, computationally we are unable to explore the model space.

Computational problems

The same issues encountered earlier are experienced, namely, computationally we are unable to explore the model space.

There are some work-arounds, involving integrating our z (these are known as continuous shrinkage priors), but we no longer explore the model space.


Q
U
E
N
S
?

Reference I

[Gir21] Christophe Giraud. *Introduction to High-Dimensional Statistics*. Chapman and Hall/CRC, Aug. 2021.

[Sim+13] Noah Simon et al. "A sparse-group lasso". In: *Journal of Computational and Graphical Statistics* 22.2 (2013), pp. 231–245.